# Computational prediction of lysine pupylation sites in prokaryotic proteins using position specific scoring matrix into bigram for feature extraction

Vineet Singh[1,✉], Alok Sharma[2,3,4,5,6,✉], Abel Chandra[5], Abdollah Dehzangi[7], Daichi Shigemizu[3,4,6,8], Tatsuhiko Tsunoda[3,4,6]

[1] Faculty of Science Technology and Environment, University of the South Pacific, Suva, Fiji
[2] Institute for Integrated and Intelligent Systems, Grifth University, Brisbane, QLD-4111, Australia
[3] Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, 113-8510, Japan
[4] Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama, 230-0045, Kanagawa, Japan
[5] School of Engineering and Physics, Faculty of Science Technology and Environment, University of the South Pacific, Suva, Fiji
[6]CREST, JST, Tokyo, 113-8510, Japan
[7]Department of Computer Science, Morgan State University, Baltimore, Maryland, USA
[8]Division of Genomic Medicine, Medical Genome Center, National Center for Geriatrics and Gerontology, Obu 474-8511, Japan
vineet.singh@usp.ac.fj, alok.sharma@griffith.edu.au

**Abstract.** Post-transcriptional modification (PTM) in a form of covalently attached proteins like ubiquitin (Ub) are considered an exclusive feature of eukaryotic organisms. Pupylation, a crucial type of PTM of prokaryotic proteins, is modification of lysine residues with a prokaryotic ubiquitin-like protein (Pup) tagging functionally to ubiquitination used by certain bacteria in order to target proteins for proteasomal degradation. Pupylation plays an important role in regulating many biological processes and accurate identification of pupylation sites contributes in understanding the molecular mechanism of pupylation. The experimental technique used in identification of pupylated lysine residues is still a costly and time-consuming process. Thus, several computational predictors have been developed based on protein sequence information to tackle this crucial issue. However, the performance of these predictors are still unsatisfactory. In this work, we propose a new predictor, PSSM-PUP that uses evolutionary information of amino acids to predict pupylated lysine residues. Each lysine residue is defined through its profile bigrams extracted from position specific scoring matrices (PSSM). PSSM-PUP has demonstrated improvement in performance compared to other existing predictors using the benchmark dataset from Pupdb Database. The proposed method achieves highest performance in 10-fold PSSM-PUP with accuracy value of 0.8975, sensitivity value of 0.8731, specificity value of 0.9222, precision value of 0.9222 and Matthews correlation coefficient value of 0.801.

**Keywords:** Post-translational modification, lysine pupylation prediction, position specific scoring matrices (PSSM) ·

## 1 Introduction

The chemical alterations of proteins after being transformed in the ribosome creates a relevant biological reaction in the cell. Post-translational modification (PTM) is alteration of amino acids in the protein sequence, which contributes to diversify the proteome [1, 2] There are many PTMs, listed from methylation [3] and ubiquitination [4] to acetylation [5], succinylation [6, 7] and phosphoglycerylation [8]. Recently, the scientific community are looking into another PTM called pupylation. A bacterial prokaryotic ubiquitin-like protein (Pup) is an intrinsically unstructured protein with 64 amino acids [9]. Pupylation is a process of Pup attaching substrate lysine via is opeptide bonds which plays important role in regulating various cellular processes such as protein degradation and signal transduction in prokaryotic cells [10]. Although pupylation and ubiquitylation are functionally the same, the enzymology involved in the two are different. Ubiquitylation requires three enzymes (activating enzyme, conjugating enzyme, and protein ligase), whereas pupylation requires only two enzymes; deamidase of Pup (DOP) and proteasome accessory factor A (PafA) [11-13]. Firstly, C-terminal glutamine of Pup is deamidated to glutamate via DOP and then deamidated Pup is attached to specific lysine of substrate proteins by PafA. The prokaryotic pupylation is still mostly unknown [14-16].

It is important to accurately identify pupylation sites to understand the fundamental mechanisms of pupylation. The traditional wet-lab experiment to identify pupylated site is expensive, inefficient and time-consuming and therefore computational tools for prediction are essential. Although there a number of computational methods developed for this, the prediction performance is still unsatisfactory. The first predictor to predict pupylation sites was proposed by Liu et al., called GPS-PUP which used a group-based prediction system (GPS) sequence encoding [17]. Zhao et al., employed the bi-profile Bayes feature extraction with support vector machine (SVM) classifier to develop EnsemblePup [18]. Zhao et al., also proposed another computational predictor PrePup which uses multiple feature encoding such as position-specific scoring matrix (PSSM) conservation scores, structural disorder score, amino acid index property (AAindex), secondary structure, solvent accessibility, and feature space with a SVM classifier [19]. Another computational predictor PUL-PUP, was established by Jiang and Cao using positive-unlabeled learning with a composition of k-spaced amino acid pairs feature (CKSAAP) and SVM algorithm [20]. Ju et al., proposed a predictor IMP-PUP by constructing features based on the composition of k-spaced amino acid pairs and on the basis of semi-supervised self-training SVM algorithm [21]. In SVM based predictor iPUP, Tung et al., also used the CKSAAP [22]. Chen et al proposed PupPred, where the sequential, structural and evolutionary hallmarks around pupylation sites were investigated and employed some of the sequence-derived features [23]. The features included physicochemical properties, binary features, protein secondary structures, amino acid pairs and PSSM with a k-nearest neighbor algorithm in SVM-based classifier. Hasan et al., developed pbPUP predictor on the basis of profile-based composition

of k-spaced amino acid pair (pbCKSAAP) encoding with SVM classifier [24]. In recent paper by Hasan. et al., shows the progress and challenges faced in protein pupylation sites prediction [25]. Most recently, Xuanguo et al., proposed an enhanced positive-unlabeled learning algorithm (EPuL) which employs only positive and unlabeled samples. The EPuL algorithm is implemented to select the reliably negative initial dataset and then iteratively picking out the non-pupylation sites [26]. In very recent work, Bao et al., developed CIPPN which identifies pupylation sites using neural network [27]. Most of the predictors have used the benchmark datasets from the PupDB database [28].

Although there are several predictors available, performance of pupylated lysine residues prediction remains unsatisfactory. Therefore, better approaches are needed by using relevant characteristics of amino acids for perception information. From the existing predictors, PrePup [19] and PupPred [23] incorporated evolutionary information, but performance can be further improved. In this work, we propose a new predictor named as PSSM-PUP (position specific scoring matrix into bigram for pupylation prediction) which employs evolutionary features of amino acids where we computed PSSM for each protein for predicting pupylated lysines. We selected a segment comprising 21 amino acids, 10 upstream and 10 downstream corresponding to each lysine residue for feature extraction. Afterward, profile bigram [29] was computed on this segment which is used to define the features of lysine residue. Since there is not enough information available from the knowledge of primary sequences, PSSM-PUP is designed to obtain information by evaluating each protein sequence related to pupylation sites.

For this work, we used a benchmark dataset consisting of 153 proteins from PupDB database [28]. This dataset has a very high number of non-pupylated lysine residues (negative samples) over the pupylated lysine residues (positive samples). We employed k-nearest neighbors cleaning treatment [1] to reduce this imbalance. Finally, a LIBSVM (library for support vector machines) package was used to develop pupylation prediction. PSSM-PUP has shown improvement in performance compared to existing predictors [20, 21].

## 2 Materials and Methods

This paper discusses the predictor called PSSM-PUP, which uses PSSM of a protein with the profile bigram of amino acids around lysines to predict pupylated and non-pupylated lysine residues [29]. The following sections discusses the benchmark data used for this study, extraction of evolutionary feature via PSSM, computation of profile bigram from PSSM for a segment of amino acids around corresponding lysine reside and SVM classifier used for pupylation prediction.

### 2.1 Benchmark Dataset

The dataset used in this study was downloaded from PupDB database [28]. It comprises of 153 protein sequences with pupylated and non-pupylated lysine residues. All

the protein sequences were used for computing the sequence identity of the dataset. We used the cd-hit program [30] to have less than 40% sequence alignment. We evaluated each protein sequence and retrieved its pupylated and non-pupylated lysine residues. We obtained 181 pupylation sites (positive samples) and 2290 non-pupylation sites (negative samples).

### 2.2 Evolutionary feature via PSSM

For a given amino acid, PSSM gives its substitution probability with the 20 amino acids of the human genome, according to its location in the protein sequence. These probabilities are obtained using PSI-BLAST tool that aligns the protein sequence to similar sequences found in the protein data bank [31]. PSI-BLAST calculates the probabilities for all the protein sequences in our benchmark dataset. The output of this tool are two L x 20 matrices in which L represents the protein sequence length and 20 the amino acids of the genetic code. One matrix is called log-odds, while the other one, the linear probabilities of amino acids. In this work, the latter is used, i.e., the linear probabilities of amino acids. PSSM extracts promising features relevant for evolutionary information [32-38].

### 2.3 Feature Extraction

In this work, PSSM feature is used to discriminate the pupylated and non-pupylated sites by considering 10 downstream and 10 upstream amino acids to the lysine residue. The lysine residue in the center, with downstream and upstream amino acids (see Fig. 1) and makes a total window size equal to 21. We computed predictor's performance with window sizes of 15, 21, 25, 27, 31, 37, 41 and 21 gave the best result. Four of the previous studies [19-21, 26]also used window size 21 for pupylation prediction. For the case where a lysine is located towards the N or C terminus of the protein sequence and there are not enough residues for either downstream or upstream, the mirroring effect [1, 6, 8, 39, 40] is used (see Fig. 2). Usage of the mirror technique to deal with the issue of insufficient residues may not be biologically correct procedure, but it has been the most effective solution by far. We can represent each lysine residue with 10 amino acids downstream and 10 amino acids upstream by

$$S = [L_{-10}, L_{-9}, ..., L_{-2}, L_{-1}, K, L_1, L_2, ..., L_9, L_{10}] \quad (1)$$

The residues $L_{-i}$ $(1 \leq i \leq 10)$ are the upstream amino acids and $L_i$ $(1 \leq i \leq 10)$ are the downstream amino acids. It can be observed from Eq. (1) that each lysine residue is represented by 21 amino acids, including the lysine itself in the center. The segment $S$ describing each lysine residue belongs to one of the two classes (c={0,1}), where a non-pupylated site falls in class 0 (c = 0) while a pupylated site is categorized as class 1 (c = 1). The vector that represents each segment $S$ are extracted from the PSSM values obtained for the entire protein sequence. Furthermore, this vector was transformed into frequency vector with bigram [29]. The resulting 20 x 20 matrix obtained after the

PSSM + bigram transformation was reordered into a 400-dimensional vector, which are the evolutionary features representing the segment *S*.

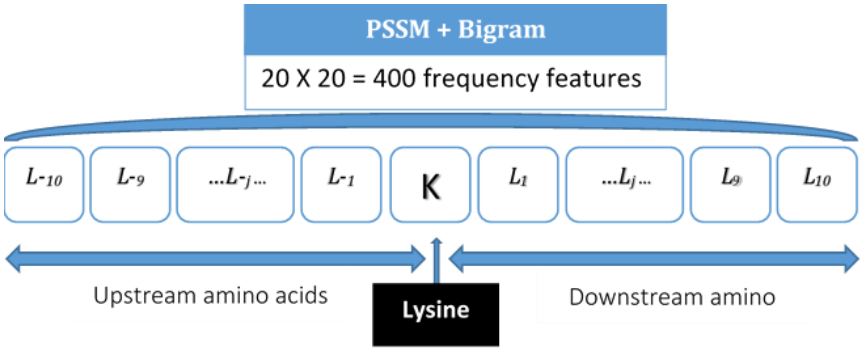


**Fig. 1.** Shows neighboring residues to the one lysine residues (K). Lysine site with enough upstream and downstream amino acids.

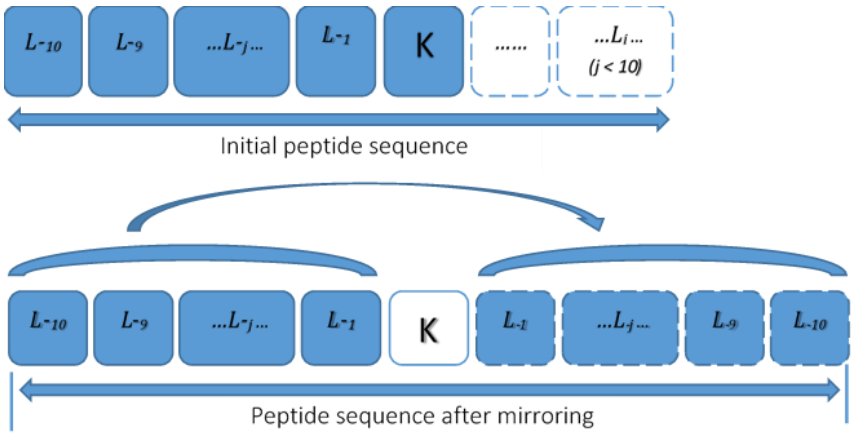


**Fig. 2.** Illustrates lysine with insufficient number of amino acids on either left or right of lysine residue (K). Left mirroring carried out to get adequate upstream and right mirroring is done to get missing downstream amino acids.

The PSSM + bigram procedure and how each segment *S* are represented is outlined below. PSSM obtained from PSI-BLAST for each protein sequence is a matrix of size L x 20. Each element of the matrix, which can be labeled as $m_{ij}$, indicates the transitional probability of *j*-th amino acid at *i*-th location in the protein sequence concerned. In this manner, PSSM results in the substitution probabilities of the 20 amino acids for the given protein sequence. The segment *S*, which is a small part of the entire protein sequence, is therefore a 21 x 20 feature vector after the extraction. The profile bigram [29] of segment size 21 was calculated by

$$B_{p,q} = \sum_{k=1}^{20} m_{k,p} m_{k+1,q} \quad \text{where } 1 \leq p \leq 20 \text{ and } 1 \leq q \leq 20 \tag{2}$$

The Eq. (2) returns 400 frequencies that correspond to 400 bigram transitions. Profile bigram is known to give good performance in the different areas of protein analysis [29, 41, 42]. The matrix $B$ (PSSM + bigram) was reordered into a 400 element feature vector $F$ as shown in Eq. (3) below. The superscript $T$ denotes transpose.

$$F = [B_{1,1}, B_{1,2}, \ldots, B_{1,20}, B_{2,1}, B_{2,2}, \ldots, B_{2,20}, B_{20,1}, B_{20,2}, \ldots, B_{20,20}]^T \tag{3}$$

The evolutionary information was computed for the 181 lysine residues in the positive set (c = 1), as well as for the 2471 in the negative set (c = 0). It is worth noting that this method provides a 400-dimensional feature vector in spite of the length of the segment size. This is an important property of profile bigram where the size of feature vector does not increase when larger segment sizes are used.

### 2.4 Support Vector Machine

SVM [43] is one type of supervised learning algorithm in the field of machine learning. SVM has been used for both regression and classification purposes but is mostly common for classification tasks and used in many existing pupylation predictors [19-22, 24, 26]. The way this algorithm works is by finding a hyperplane that best discriminates the two classes i.e. it finds a plane that has the maximum distance between data points of the two classes. Moreover, the number of features of these data points has the effect on the dimensionality of the hyperplane. For instance, feature size of 2 requires a hyperplane that is 1 dimensional (a line). Furthermore, not all classes are linearly separable. In these cases, non-linear kernels are used. Non-linear kernels map the nonlinear input space to a feature space of higher dimension in which the classes can be linearly separated. LIBSVM [44] predictor has been employed in this work on Matlab platform and the SVM type selected was radial basis function kernel and cost value of 2 and gamma value of 0.0250.

### 2.5 Statistical measures

To evaluate the performance of the proposed predictor and compare with the existing predictors, few measures which are sensitivity (*Sn*), specificity (*Sp*), accuracy (*Acc*), precision (*Pre*) and Matthews correlation coefficient (*MCC*) are employed in this work.

One of the key measure is sensitivity, which evaluates the percentage of pupylated residues correctly classified by the model. The predictor achieving high sensitivity shows that it can accurately detect those positive instances (pupylated residues) in the dataset. Simply when sensitivity equals to 1 makes an accurate predictor and when it equals to 0 makes it an inaccurate one. The formula for sensitivity is defined as:

$$Sensitivity = \frac{PL_+}{PL_+ + PL_-} \tag{4}$$

where $PL_+$ is number of pupylated lysine predicted correctly and $PL_-$ represents the number of pupylated lysine incorrectly classified by the predictor

On the other hand, specificity assesses the proportion of correctly identified non-pupylated lysine residues. Specificity of 1 demonstrates an accurate predictor which is able to predict negative instance of the dataset (non-pupylated residues) and specificity equals to 0 shows predictor is unable to identify non-pupylated residues. The metric for specificity is defined as

$$Specificity = \frac{NPL_+}{NPL_+ + NPL_-} \quad (5)$$

where $NPL_+$ is the number of non-pupylated lysine predicted correctly and $NPL_-$ represents the number of incorrectly classified non-pupylated lysine by the predictor

For a predictor to correctly distinguish between positive samples and negative samples is evaluated by the accuracy of the predictor. Predictor with accuracy equals to 1 shows an accurate predictor whereas a zero accuracy means predictor is totally incorrect. Accuracy is calculated as

$$Accuracy = \frac{PL_+ + NPL_+}{PL + NPL} \quad (6)$$

where $PL$ and $NPL$ are the total numbers of pupylated and non-pupylated lysine residues, respectively.

Precision is another assessment measure of the predictor defined as the ratio of the number of correctly identify pupylated lysine over sum of correctly classified pupylated and non-pupylated lysine residues.

$$Precision = \frac{PL_+}{PL_+ + NPL_+} \quad (7)$$

Final statistical measure used in this paper is the Matthews correlation coefficient (MCC). It shows the value of correlation coefficient between predicted and observed instances. If a predictor has MCC equals to 1, it implies a perfect correlation between prediction and observation whereas, MCC equals to -1 does not show any agreement. MCC metric is calculated as

$$MCC = \frac{(NPL_+ \times PL_+) - (NPL_- \times PL_-)}{\sqrt{(PL_+ + PL_-)(PL_+ + NPL_-)(NPL_- + PL_-)(NPL_+ + NPL_-)}} \quad (8)$$

A best predictor is the one that achieves high performance in the five statistical measures discussed. However, it should perform better at least in some of the measures compared to the existing predictors. A predictor which is unable to predict pupylated lysine correctly (low sensitivity) cannot be used for pupylation prediction.

### 2.6 Validation Scheme

The effectiveness of a new predictor needs to be assessed with a validation method. There are several validation methods discussed in literature, however, two most used ones are the jackknife and n-fold validation scheme [45, 46]. In validation phase, an independent test set has to be used to assess the predictor. The Jackknife validation is less arbitrary than the n-fold cross-validation and provides unique results for a dataset [47]. From the literature, the same validation scheme [19-22, 26, 48] (n-fold cross-validation) technique is used in this study. The n-fold cross-validation technique is carried out in following steps listed in table 1:

**Table 1.** Steps for cross-validation approach

1. Split the data samples complementary into n folds of roughly equal sample size with similar positive and negative sample size in each.
2. Use one fold as independent test set and the remaining $n-1$ folds as training data.
3. Use the training data, adjust the parameters of the predictor
4. Compute all the statistical measures on independent test set
5. Repeat steps 1 to 4 for the remaining folds for assessment and calculated the average of each statistical measure.

In this study, we conducted 6-, 8- and 10-fold cross-validations for assessing the PSSM-PUP predictor and result were recorded

## 3 Results and Discussion

Any proposed predictors need to be assessed in order to measure it performance. For this study, we used five statistical metrics: sensitivity, specificity, precision, accuracy and Matthews correlation coefficient [19, 20, 22, 24, 49] which are commonly used in the literature. The following sections discusses how the class imbalance were treated and also presents the results of support vector machine classification. The overall performance of PSSM-PUP and comparison with existing pupylation predictors with five metrics are also discussed.

### 3.1 Reducing the imbalance between classes

After analyzing the protein sequence of our dataset, we found out the number of positive samples (pupylation sites) is much smaller than the negative samples (non-pupylation sites). This led to a high class imbalance samples that can cause biased classification results. Imbalance between samples of different classes is a common issue in machine learning and it is crucial to mitigate this problem. This proposed predictor removes redundant instances before the classification takes place. We used k-nearest neighbor technique in this study to deal imbalance of samples between classes. K-nearest neighbor technique is very popular in pattern recognition which was reintroduced for protein attribute prediction by Chou [50]. To balance both negative and positive classes, we removed redundant negative samples using k-nearest neighbors cleaning treatment [25]. We calculated Euclidean distance between all the samples in the dataset. We first set the cut-off by dividing the number of negative instance and positive instances (2,290/181) which came to a ratio of 12.65. Thus, K=12 was initially set for reducing class imbalance. In other terms, we remove a negative sample if one of 12 nearest neighbor is a positive sample (calculation based on the Euclidean distance between the negative sample and all other samples in the entire dataset). After this first filtering, the imbalance classes still remained, therefore, we kept increasing the K value

until the both the sets were almost similar in size. This method reduced the initial negative samples of 2,290 to 180 with a threshold value of 70, meaning a negative sample was removed if at least one positive sample is present within the 70 nearest neighbor. The negative instances were reduced to 180 samples. The positive instances remained 181 as it can affect the sensitivity. The final dataset after filtering (filtered negative samples and positive samples) was used to carry out 6-, 8-, 10- fold cross-validation and assess the predictor's performance.

### 3.2 Comparison with existing predictors

We compared our proposed PSSM-PUP predictor with two recently proposed predictors: PuL-PUP [19] by Jiang and Cao, and IMP-PUP [20] by Ju et al. Unfortunately, we could not compare with EPuL algorithm [25] since the given webserver was not working and the software package also did not work. The software package for testing were given for these two predictors PuL-PUP [19] and IMP-PUP [20]. Since all exiting predictors used the same dataset, it is worth noting that the trained model in existing predictors would have utilized some of the same protein sequences in their training which are in my test samples. Therefore, for comparison purposes, we used the feature extraction method to extract the features from the given software package and trained and tested using the LIBSVM classifier. The same train and test sets used in our proposed PSSM-PUP predictor was used to train and test for different folds when comparing with other predictors. We calculated the sensitivity, specificity, precision, accuracy, MCC for PSSM-PUP, PuL-PUP and IMP-PUP for 6-, 8- and 10-fold cross-validation trials.

**Table 2.** Table shows performance assessment of two benchmark predictors and PSSM-PUP for 6-, 8-, 10- fold cross validation. The highest values in each metric are highlighted in bold.

| Fold | Predictor | Sensitivity | Specificity | Precision | Accuracy | MCC |
|---|---|---|---|---|---|---|
| 6 | PSSM-PUP | 85.645 | 92.222 | 91.920 | 88.916 | 0.782 |
| | PUL-PUP | 80.054 | 74.444 | 76.188 | 77.272 | 0.552 |
| | IMP-PUP | 82.276 | 72.222 | 74.856 | 77.272 | 0.549 |
| 8 | PSSM-PUP | 85.598 | 92.762 | 92.523 | 89.174 | 0.788 |
| | PUL-PUP | 79.891 | 77.841 | 78.552 | 78.873 | 0.583 |
| | IMP-PUP | 81.719 | 72.826 | 75.231 | 77.280 | 0.551 |
| 10 | PSSM-PUP | **87.310** | **92.222** | **92.290** | **89.752** | **0.801** |
| | PUL-PUP | 81.754 | 76.111 | 77.693 | 78.956 | 0.584 |
| | IMP-PUP | 82.310 | 70.556 | 74.145 | 76.441 | 0.537 |

The comparison of predictor PuL-Pup [19], IMP-PUP [20] with PSSM-PUP is shown Table 1. Improvement in performance for PSSM-PUP is seen over PuL-Pup [19] and IMP-PUP [20] on sensitivity, specificity, precision, accuracy and MCC. The performance improved slightly for sensitivity but significantly for specificity, precision,

accuracy and MCC. It is worth noting that only the feature extraction methods were used to get the training and test sets. Same LIBSVM classifier with the same SVM parameters was used to train and test the predictors for comparison.

The promising results shows the ability of proposed PSSM-PUP predictor to correctly identify pupylated and non-pupylated lysine residues. This is possible since proposed predictor uses significant evolutionary information of protein sequences effectively. This information which is stored in the PSSM of each amino acid around lysine, when placed in one matrix of bigram shows important characteristic for detecting modified lysines. The SVM classifier and its effective use in PTM also improves the outcome. In short, the combination of PSSM + bigram extracts more information around lysine residues, which plays a vital role in predicting pupylated and non-pupylated lysine residues.

Our PSSM-PUP predictor's software package can be accessed from: https://github.com/vinzsingh09/PSSM-PUP.

## 4 Conclusion

This paper discussed a new predictor named PSSM-PUP, which has used the combination of *PSSM + Bigram* efficiently for pupylation prediction. The evolutionary information hidden in PSSMs that is converted to bigram occurrences shows to be a significant feature which can used for prediction. The k-nearest neighbors cleaning treatment also plays an important role to solve imbalance data issue and removing redundant samples to balance the dataset. A balanced dataset with support vector machine (LIBSVM) has shown PSSM-PUP to perform better than exiting existing predictors. For future study, we intend to use structural properties of amino acids for pupylation prediction and further explore the use of a 21-residue window for describing lysine residues.

## References

1. Jia, J., Liu, Z., Xiao, X., Liu, B., Chou, K.-C.: iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. Analytical biochemistry 497, 48-56 (2016)
2. Walsh, C.T., Garneau-Tsodikova, S., Gatto Jr, G.J.: Protein posttranslational modifications: the chemistry of proteome diversifications. Angewandte Chemie International Edition 44, 7342-7372 (2005)
3. Liu, Z., Xiao, X., Qiu, W.-R., Chou, K.-C.: iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition. Analytical biochemistry 474, 69-77 (2015)
4. Qiu, W.-R., Xiao, X., Lin, W.-Z., Chou, K.-C.: iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model. Journal of Biomolecular Structure and Dynamics 33, 1731-1742 (2015)

5. Hou, T., Zheng, G., Zhang, P., Jia, J., Li, J., Xie, L., Wei, C., Li, Y.: LAceP: lysine acetylation site prediction using logistic regression classifiers. PloS one 9, e89575 (2014)
6. Dehzangi, A., López, Y., Lal, S.P., Taherzadeh, G., Michaelson, J., Sattar, A., Tsunoda, T., Sharma, A.: PSSM-Suc: Accurately predicting succinylation using position specific scoring matrix into bigram for feature extraction. Journal of theoretical biology 425, 97-102 (2017)
7. López, Y., Dehzangi, A., Lal, S.P., Taherzadeh, G., Michaelson, J., Sattar, A., Tsunoda, T., Sharma, A.: SucStruct: Prediction of succinylated lysine residues by using structural properties of amino acids. Analytical biochemistry 527, 24-32 (2017)
8. Chandra, A., Sharma, A., Dehzangi, A., Ranganathan, S., Jokhan, A., Chou, K.-C., Tsunoda, T.: PhoglyStruct: Prediction of phosphoglycerylated lysine residues using structural properties of amino acids. Scientific reports 8, 17923 (2018)
9. Burns, K.E., Liu, W.-T., Boshoff, H.I., Dorrestein, P.C., Barry, C.E.: Proteasomal protein degradation in Mycobacteria is dependent upon a prokaryotic ubiquitin-like protein. Journal of Biological Chemistry 284, 3069-3075 (2009)
10. Chen, X., Solomon, W.C., Kang, Y., Cerda-Maira, F., Darwin, K.H., Walters, K.J.: Prokaryotic ubiquitin-like protein pup is intrinsically disordered. Journal of molecular biology 392, 208-217 (2009)
11. Burns, K.E., Cerda-Maira, F.A., Wang, T., Li, H., Bishai, W.R., Darwin, K.H.: "Depupylation" of prokaryotic ubiquitin-like protein from mycobacterial proteasome substrates. Molecular cell 39, 821-827 (2010)
12. Imkamp, F., Striebel, F., Sutter, M., Özcelik, D., Zimmermann, N., Sander, P., Weber-Ban, E.: Dop functions as a depupylase in the prokaryotic ubiquitin-like modification pathway. EMBO reports 11, 791-797 (2010)
13. Striebel, F., Imkamp, F., Özcelik, D., Weber-Ban, E.: Pupylation as a signal for proteasomal degradation in bacteria. Biochimica et Biophysica Acta (BBA)-Molecular Cell Research 1843, 103-113 (2014)
14. Striebel, F., Imkamp, F., Sutter, M., Steiner, M., Mamedov, A., Weber-Ban, E.: Bacterial ubiquitin-like modifier Pup is deamidated and conjugated to substrates by distinct but homologous enzymes. Nature structural & molecular biology 16, 647 (2009)
15. Georgiou, D., Karakasidis, T., Megaritis, A.: A short survey on genetic sequences, Chou's pseudo amino acid composition and its combination with fuzzy set theory. Open Bioinform. J 7, 41-48 (2013)
16. Poulsen, C., Akhter, Y., Jeon, A.H.W., Schmitt-Ulms, G., Meyer, H.E., Stefanski, A., Stühler, K., Wilmanns, M., Song, Y.H.: Proteome-wide identification of mycobacterial pupylation targets. Molecular systems biology 6, 386 (2010)
17. Liu, Z., Ma, Q., Cao, J., Gao, X., Ren, J., Xue, Y.: GPS-PUP: computational prediction of pupylation sites in prokaryotic proteins. Molecular BioSystems 7, 2737-2740 (2011)
18. Zhao, X., Zhang, J., Ning, Q., Sun, P., Ma, Z., Yin, M.: Identification of protein pupylation sites using bi-profile Bayes feature extraction and ensemble learning. Mathematical Problems in Engineering 2013, (2013)
19. Zhao, X., Dai, J., Ning, Q., Ma, Z., Yin, M., Sun, P.: Position-specific analysis and prediction of protein pupylation sites based on multiple features. BioMed research international 2013, (2013)
20. Jiang, M., Cao, J.-Z.: Positive-Unlabeled learning for pupylation sites prediction. BioMed research international 2016, (2016)

21.Ju, Z., Gu, H.: Predicting pupylation sites in prokaryotic proteins using semi-supervised self-training support vector machine algorithm. Analytical biochemistry 507, 1-6 (2016)
22.Tung, C.-W.: Prediction of pupylation sites using the composition of k-spaced amino acid pairs. Journal of theoretical biology 336, 11-17 (2013)
23.Chen, X., Qiu, J.-D., Shi, S.-P., Suo, S.-B., Liang, R.-P.: Systematic analysis and prediction of pupylation sites in prokaryotic proteins. PloS one 8, e74002 (2013)
24.Hasan, M.M., Zhou, Y., Lu, X., Li, J., Song, J., Zhang, Z.: Computational identification of protein pupylation sites by using profile-based composition of k-spaced amino acid pairs. PloS one 10, e0129635 (2015)
25.Hasan, M.M., Khatun, M.S.: Recent progress and challenges for protein pupylation sites prediction. EC Proteomics and Bioinformatics 2, 36-45 (2017)
26.Nan, X., Bao, L., Zhao, X., Zhao, X., Sangaiah, A., Wang, G.-G., Ma, Z.: EPuL: an enhanced positive-unlabeled learning algorithm for the prediction of pupylation sites. Molecules 22, 1463 (2017)
27.Bao, W., You, Z.-H., Huang, D.-S.: CIPPN: computational identification of protein pupylation sites by using neural network. Oncotarget 8, 108867 (2017)
28.Tung, C.-W.: PupDB: a database of pupylated proteins. BMC bioinformatics 13, 40 (2012)
29.Sharma, A., Lyons, J., Dehzangi, A., Paliwal, K.K.: A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. Journal of theoretical biology 320, 41-46 (2013)
30.Li, W., Godzik, A.: Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658-1659 (2006)
31.Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., Bourne, P.: The Protein Data Bank Nucleic Acids Research, 28, 235-242. URL: www. rcsb. org Citation (2000)
32.Dehzangi, A., Paliwal, K., Lyons, J., Sharma, A., Sattar, A.: Proposing a highly accurate protein structural class predictor using segmentation-based features. BMC genomics 15, S2 (2014)
33.Faraggi, E., Zhang, T., Yang, Y., Kurgan, L., Zhou, Y.: SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. Journal of computational chemistry 33, 259-267 (2012)
34.Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., Sattar, A., Yang, Y., Zhou, Y.: Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. Scientific reports 5, 11476 (2015)
35.McGuffin, L.J., Bryson, K., Jones, D.T.: The PSIPRED protein structure prediction server. Bioinformatics 16, 404-405 (2000)
36.Taherzadeh, G., Zhou, Y., Liew, A.W.-C., Yang, Y.: Sequence-based prediction of protein–carbohydrate binding sites using support vector machines. Journal of chemical information and modeling 56, 2115-2122 (2016)
37.Taherzadeh, G., Yang, Y., Zhang, T., Liew, A.W.C., Zhou, Y.: Sequence-based prediction of protein–peptide binding sites using support vector machine. Journal of computational chemistry 37, 1223-1229 (2016)

38.Dehzangi, A., Paliwal, K., Lyons, J., Sharma, A., Sattar, A.: A segmentation-based method to extract structural and evolutionary features for protein fold recognition. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 11, 510-519 (2014)
39.Jia, J., Liu, Z., Xiao, X., Liu, B., Chou, K.-C.: pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. Journal of theoretical biology 394, 223-230 (2016)
40.López, Y., Sharma, A., Dehzangi, A., Lal, S.P., Taherzadeh, G., Sattar, A., Tsunoda, T.: Success: evolutionary and structural properties of amino acids prove effective for succinylation site prediction. BMC Genomics 19, 923 (2018)
41.Dehzangi, A., Heffernan, R., Sharma, A., Lyons, J., Paliwal, K., Sattar, A.: Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou′ s general PseAAC. Journal of theoretical biology 364, 284-294 (2015)
42.Dehzangi, A., López, Y., Lal, S.P., Taherzadeh, G., Sattar, A., Tsunoda, T., Sharma, A.: Improving succinylation prediction accuracy by incorporating the secondary structure via helix, strand and coil, and evolutionary information from profile bigrams. PloS one 13, e0191900 (2018)
43.Meyer, D., Leisch, F., Hornik, K.: Benchmarking support vector machines. (2002)
44.Chang, C.-C.: " LIBSVM: a library for support vector machines," ACM Transactions on Intelligent Systems and Technology, 2: 27: 1--27: 27, 2011. http://www. csie. ntu. edu. tw/~ cjlin/libsvm 2, (2011)
45.Chou, K.-C., Shen, H.-B.: Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nature protocols 3, 153 (2008)
46.Alpaydin, E.: Introduction to machine learning. MIT press (2014)
47.Hajisharifi, Z., Piryaiee, M., Beigi, M.M., Behbahani, M., Mohabatkar, H.: Predicting anticancer peptides with Chou′ s pseudo amino acid composition and investigating their mutagenicity via Ames test. Journal of Theoretical Biology 341, 34-40 (2014)
48.Zhao, X., Ning, Q., Chai, H., Ma, Z.: Accurate in silico identification of protein succinylation sites using an iterative semi-supervised learning technique. Journal of theoretical biology 374, 60-65 (2015)
49.Bao, W., Jiang, Z.: Prediction of Lysine Pupylation Sites with Machine Learning Methods. In: International Conference on Intelligent Computing, pp. 408-417. Springer, (Year)
50.Chou, K.-C.: Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of theoretical biology 273, 236-247 (2011)